2

Model_Inference

Have the courage to follow your heart and intuition. They somehow already know what you truly want to become. Everything else is secondary.

–Steve Jobs

Data & Model

Data

열 column

목표 target 변수 Y

- 결과(output)로 나타나는, 예측을 하고픈, 관심 변수 : 반응변수, **목표변수**
- 목표변수가 2개 이상인 Simultaneous Equation Model 은 본 교과서에서 다루지 않음 - 목표변수 개수는 1개

예측 predictor 변수 $X_1, X_2, \ldots, X_p$

- 목표변수를 설명하거나, 값의 변동을 예측하는, 원인이 되는 변수 : 요인, **예측변수, Feature 변수**
- 예측변수의 개수 : p

행 , X_{ik}, X_{tk}

개체 subject Y_i - X_{ik}

- 관심의 대상이며 표본 관측치를 얻은 개체(사람, 국가)
- 시점이 고정되어 있어 이를 횡단 cross-sectional 자료

시간 time Y_t - X_{tk}

- 시계열 데이터, 시간의 순서를 갖는다.

개체와 시간의 혼용

- 데이터는 행렬의 형태로 저장되므로 횡단자료와 시계열 자료가 동시에 있는 경우는 열, 혹은 행 중 하나를 시간에 할애해야 한다.
- 예를 들어 OECD 국가 실업율, 경제성장률, GDP 자료를 최근 10년간 수집하고자 한다면, 행=35개 국가, 열=(실업율, 경제성장률, GDP), 그럼 연도는??? 열을 사용하면 되지만 일반적으로 행을 사용하는 것이 자료 다루기에 편리하다.

Model

회귀모형

예측변수 개수 p, 표본크기 n : $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + + \beta_p x_{ip} + e_i$

- α : 회귀계수(regression coefficient) 절편(intercept), 모든 x 값이 0 일때 y의 값
- $\beta_1, \beta_2, \ldots, \beta_p$: 회귀계수, 각 예측변수 X의 기울기, 예측변수 X가 한 단위 증가할 때마다 목표변수 Y의 증가량(편미분 계수) - 함수 형태 중 선형을 선호하는 이유 + 비선형 모형도 로그 변환을 통하여 선형 모형이 가능함(Cobb-Douglas 생산함수 $Q_i = \alpha K_i^{\beta} L_i^{\lambda} e_i$-$ln(Q_i) = ln(\alpha) + \beta ln(K_i) + \lambda ln(L_i) + ln(e_i)$)
- $X's$: 예측변수이며 결정변수 deterministic (즉 확률변수가 아님)
- e_i : 오차항(error term), 회귀모형에 설정된 $X's$ 에 의해 설명되지 못하는 부분, 오차항이 없으면 통계 모형이 아니라 수학모형 (모형 내의 예측변수에 의해 종변속수 값을 정확하게 예측할 수 있음)
- 즉 목표변수 값의 패턴(변동)은 회귀모형이 설명하는 변동과 모형이 설명하지 못하는 변동(수학함수이면 없음) 나누어 변동의 비를 이용하여 모형의 유의성을 검정한다.

행렬 $\underline{y} = X\underline{b} + \underline{e}$

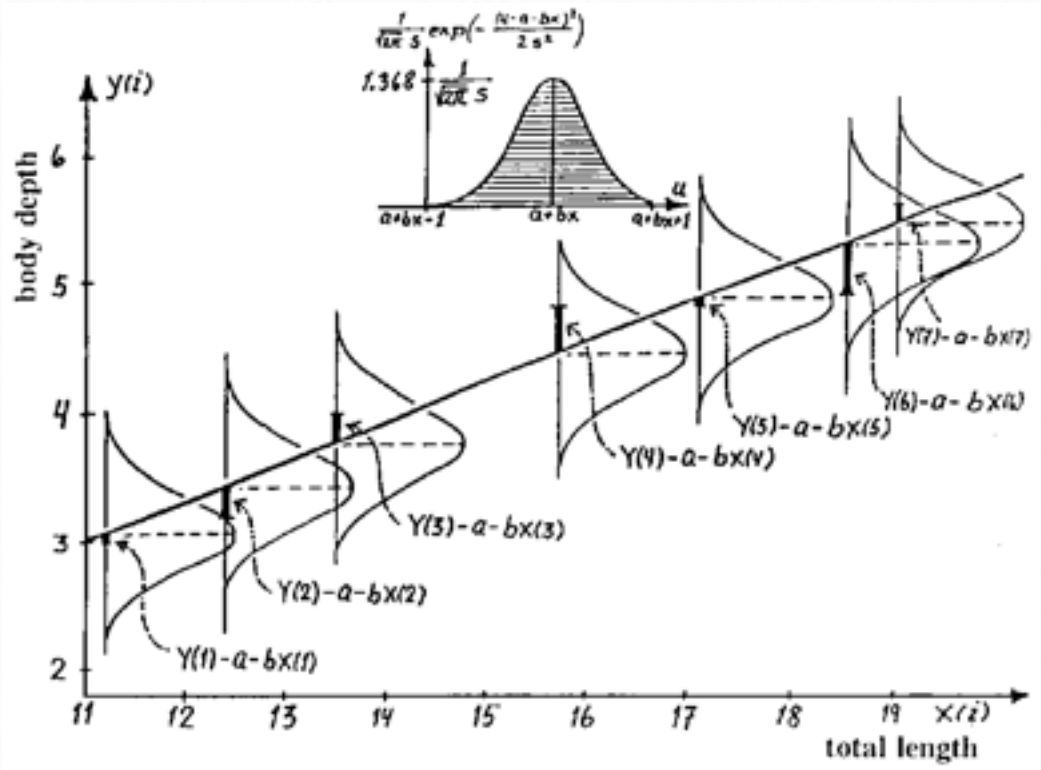
$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

(가정) $e_i \sim N(0,\sigma^2)$, $\underline{e} \sim MN(\underline{0}, \sigma^2 I)$

- 모수인 회귀계수 $\alpha, \beta_1, \beta_2, \ldots$ 는 미지이지만 하나의 상수 값 - 확률변수 아님
- 예측변수는 결정변수로 오차없이 측정할 수 있음 - 확률변수 아님
- 선형성 linearity : 목표변수와 예측변수 간 함수관계는 선형 :
- 오차항 가정 , $e_i \sim \ iidN(0,\sigma^2)$: 독립성, 정규성, 등분산성

가정 필요이유

- 독립성(independent): 오차항은 서로 독립이다. 즉 각 오차는 서로 영향을 주지 않는다. 독립성 가정은 시계열 데이터(시간적 순서를 갖는 데이터) 경우에만 체크한다.
- 정규성(normality): 오차항은 정규 분포를 따른다. 이 가정은 F-검정 방법을 사용하기 위하여 반드시 필요하다. 오차항이 정규분포를 따르므로 목표변수도 정규분포를 따르고 목표변수의 평균은 $X\underline{b}$ (선형식)
- 등분산성(Homoscedasticity): 오차항의 분산은 동일. 분산이 일정하다는 가정의 주어진 예측변수 값에서 관측되는 Y 의 값의 분산이 일정하다는 의미와 같다. 분산이 다르면 설정된 회귀 모형이 적절함에도 불구하고 관측치가 직선에 모여 있지 않게 된다. 분산이 큰 예측변수에서는 직선을 벗어나는 경향이 있어 분산이 동일하지 않은 경우에는 직선에서 벗어난 관측치가 직선의 경향을 보이지 않은 이상 관측치인지 분산의 크기가 달라 이런 현상이 타나났는지 판단할 수 없다.



분산분석 모형

요인 A(기호 i-번째 처리 수준, n_i i-수준 반복수), 요인 B(기호 j-번째 처리 수준, n_j j-수준 반복수),

교호작용 존재, k 반복 : $y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$

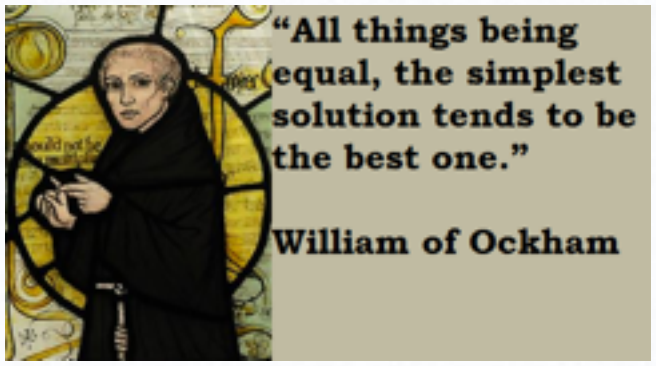
- μ : 총 평균, 목표변수의 평균으로 추정
- A_i : 요인 A i-수준의 평균 / B_j : 요인 B j-수준의 평균 / AB_{ij} : 요인 A i-수준, 요인 B j-수준의 평균
- e_{ijk} : 총변동 중 요인(A, B, AB) 변동이 설명하지 못한 변동

(가정) $e_{ijk} \sim N(0,\sigma^2)$: 독립성, 정규성, 등분산성

선형모형 분석 목표

목표변수에 유의한 영향을 주는 예측변수선택

- 오컴의 면도날(Occam's Razor) : 경제성의 원리(Principle of economy), 절약성의 원리(Principle of parsimony) - 모든 상황이 동일하다면 가장 간단한 방법이 최선이다. 즉, 회귀모형이 동일한 정보를 준다면(설정 모형이 예측변수의 변동을 설명하는 비율) 적은 예측변수 모형이 최선의 모형이다.



- 그러나 목표변수의 변동을 설명하는데 유의하지 않은 예측변수도 조금의 자신 비중(portion)은 있다. 데이터 수집의 비용이 거의 없는 빅데이터 시대에서 예측모형에서는 유의한 예측변수 선택의 필요성은 낮아진다. - 단지 목표변수에 유의한 영향을 주는 예측변수를 찾는데 사용된다.

영향을 가장 많이 미치는 예측변수 선택

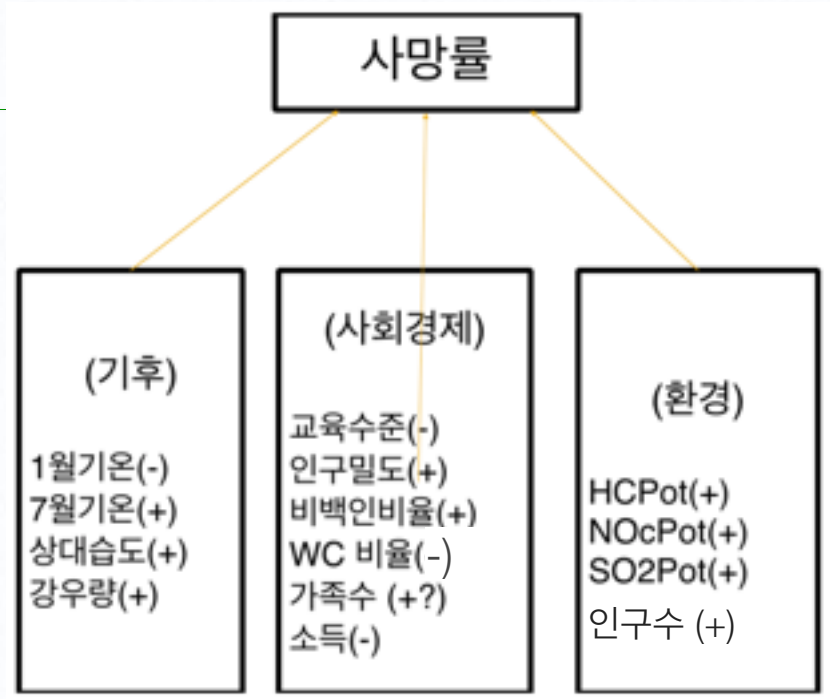
- 단위를 고려하여 표준화 회귀계수의 추정값으로 예측변수의 영향도 판단한다.

예측변수 값이 주어진 경우 목표변수 값 예측 $y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + + \hat{\beta}_p x_{ip}$

예제 데이터

변수	변수이름	변수내용
종속변수	Mortality	사망률
기후	JanTemp	1월기온
	JulyTemp	7월기온
	RelHum	상대습도
	Rain	강우량
사회경제	Education	교육수준
	PopDensity	인구밀도
	NonWhite	비백인비율
	WC	화이트칼라 비율
	pop/house	가구당 가족수
	income	소득
환경	HCPot	오염물질1
	NOxPot	오염물질2
	S02Pot	오염물질3

population. 인구수(+)



괄호 안 부호는 사망율에 미치는 영향(상관계수) 부호를 나타낸 것임 (예) 1월기온(-)의 의미는 1월 기온이 낮아질수록 사망율은 높아진다는 의미, 비백인비율(+) 의미는 비백인 비율 많은 도시의 사망율은 높을 것이라는 사전 판단 내용 [연구내용임]

```
import pandas as pd
smsa=pd.read_csv('http://203.247.53.31/Stat_Notes/example_data/SMSA_USA.csv')
smsa[['city_nm','state_nm','tmp_nm']]=smsa['city_name'].str.split(',',expand=True)
smsa.loc[smsa.tmp_nm.isna()==False, 'state_nm']=smsa.tmp_nm
smsa.drop(columns=['city_name','tmp_nm'],inplace=True)
smsa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59 entries, 0 to 58
Data columns (total 18 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   jan_temp           59 non-null     int64
 1   july_temp          59 non-null     int64
 2   humidity           59 non-null     int64
 3   rainfall           59 non-null     int64
 4   mortality_index    59 non-null     float64
 5   education          59 non-null     float64
 6   pop_density        59 non-null     int64
 7   non_white_ratio    59 non-null     float64
 8   white_color_ratio  59 non-null     float64
 9   population         59 non-null     int64
 10  person_houshold    59 non-null     float64
 11  household_income   59 non-null     int64
 12  HCPot              59 non-null     int64
 13  NOxPot             59 non-null     int64
 14  SO2Pot             59 non-null     int64
 15  southern           59 non-null     object
 16  city_nm            59 non-null     object
 17  state_nm           50 non-null     object
dtypes: float64(5), int64(10), object(3)
memory usage: 8.4+ KB
```

```
#결측치가 있는 경우 이를 제외하고 분석하는 것이 적절하다.
smsa.isna().sum()
#만약 존재한다면
smsa.dropna(inplace=True)
```

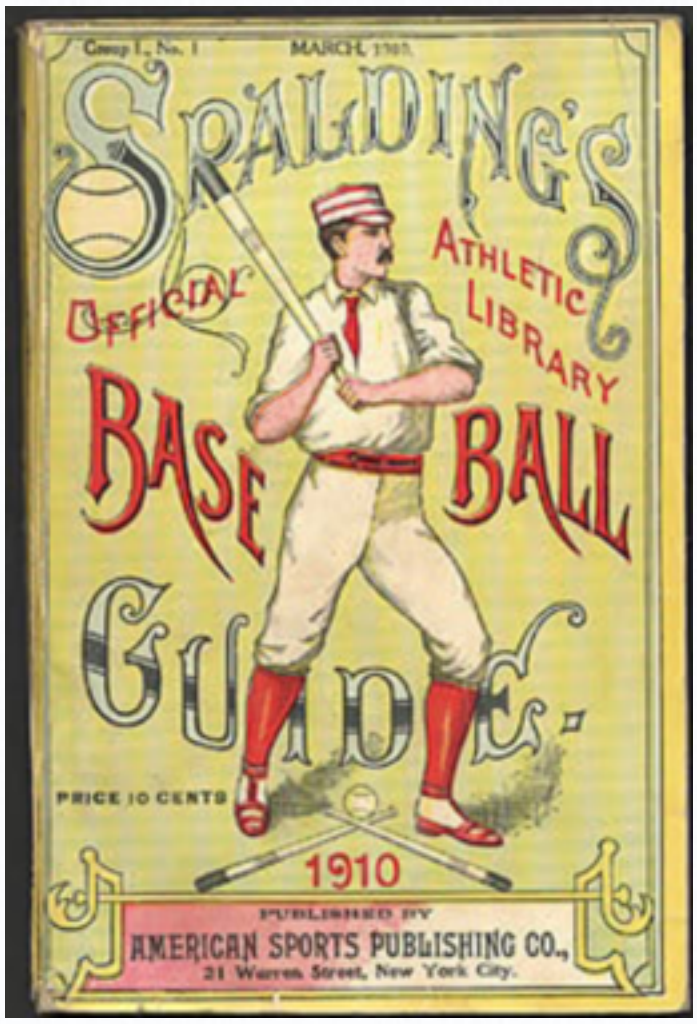
n=59 도시, 열변수=18개

Practice Data [MLB 데이터]

[연구문제] 각 구단은 타자/투수 선수 연봉 책정을 어떻게 할 것인가?

타자, 투수 http://www.seanlahman.com/baseball-archive/statistics

```
Comma Delimited Version:
      readme.txt
      AllStarFull.csv
      Appearances.csv
      AwardsManagers.csv
      AwardsPlayers.csv
      AwardsShareManagers.csv
      AwardsSharePlayers.csv
      Batting.csv
      BattingPost.csv
      CollegePlaying.csv
      Fielding.csv
      FieldingOF.csv
      FieldingPost.csv
          FieldingOFsplit
      HallOfFame.csv
          HomeGames.csv
      Managers.csv
      ManagersHalf.csv
          Parks.csv
      People.csv
      Pitching.csv
      PitchingPost.csv
          README.txt
      Salaries.csv
      Schools.csv
      SeriesPost.csv
      Teams.csv
      TeamsFranchises.csv
      TeamsHalf.csv
```




```
1  import pandas as pd
2  salary=pd.read_csv('http://203.247.53.31/Big_Data/data/MLB1871_2018/Salaries.csv')
3  salary.head(3)
```

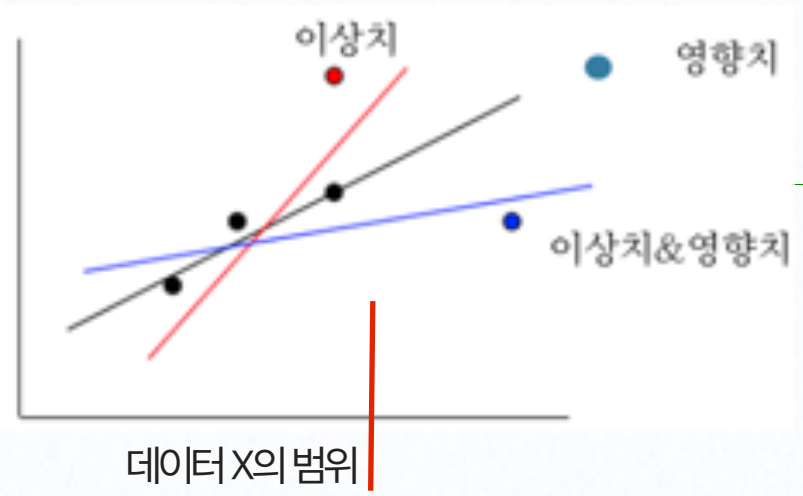
	yearID	teamID	lgID	playerID	salary
0	1985	ATL	NL	barkele01	870000
1	1985	ATL	NL	bedrost01	550000
2	1985	ATL	NL	benedbr01	545000

Scatter Plot

정의

2개의 측정형 변수 데이터를 2차원 공간에 표현하여 두 변수의 함수 관계를 예상함

- X- 축 : 결정의 요인, 예측변수, 독립변수, 예측변수
- Y-축 : Output, 목표변수, 목표변수



진단내용

- 두 변수 간의 함수 관계를 본다.
- 이상치와 영향치를 시각적으로 진단한다. 통계량을 이용한 진단은 최총모형 도출 후 회귀진단에서 실시한다.

이상치 outlier

- 선형 함수 관계에서 적합 직선을 많이 벗어난 관측값 - 실제 오차의 분산 기준 2* σ를 벗어남
- 예측변수 값은 관측 값의 범위 내에 있음

(진단) 오차의 추정치인 Studentized 잔차가 ± 2 벗어남 - 상세한 내용은 잔차 진단 참고

(해결) 삭제 - 물론 잔차분석 후에 실기

영향치 influential

- 예측변수 값이 극단 값(다른 관측치와 떨어져 있고 두 변수의 함수 관계에 영향을 주는 관측값
- 순수 영향치 : 함수 회귀 추정 식 상에 있어 함수 관계(기울기 변동)에는 영향을 주지 않으나 결정계수 높여 예측변수의 설명 능력을 과다하게 높은 것으로 판단하게 하는 결과 왜곡
- 이상치&영향치 : 결정계수 왜곡, 함수관계 왜곡

(진단) 잔차분석 - Hat 통계량 활용

(해결) 영향치 주변의 관측값을 추가 수집 후 분석, 영향치 값이 실제 발생 가능하지 않은 경우 제외

변수 분포

선형모형의 변수들은 좌우 대칭(정규분포, 특히 목표변수의 경우 오차의 정규성 가정으로 정규분포를 가져야 함)의 분포를 갖는 경우 선형 모형의 적합성과 결과 활용도가 높아짐

예전에는 각 예측변수의 분포를 산점도와 개별적으로 분석하였지만 소프트웨어의 발달로 산점도에 분포를 함께 나타낼 수 있음

코드 : 산점도 행렬

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(smsa,kind='reg',diag_kind='kde')
plt.show()
```

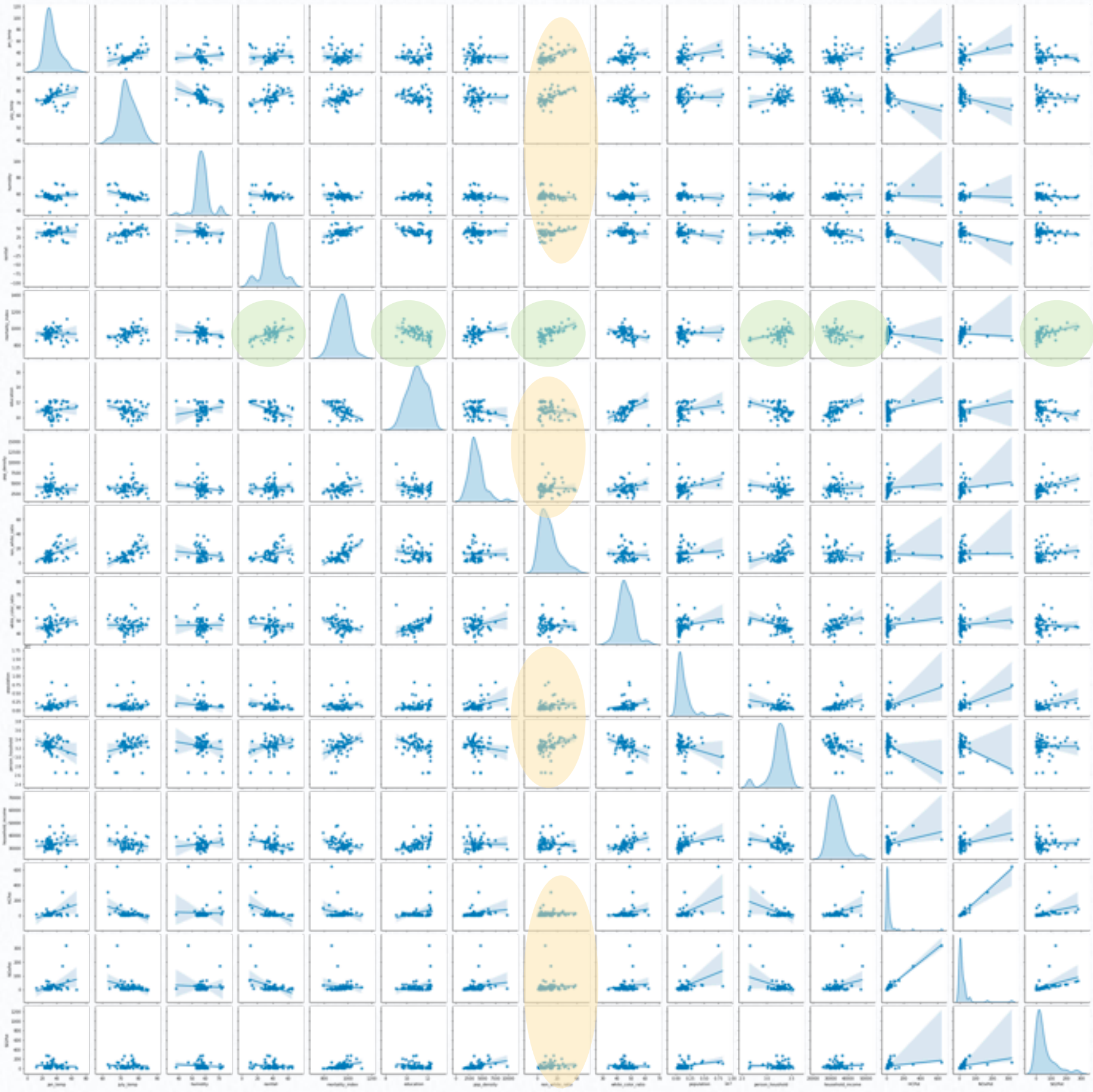
시각적 판단

종속변수와 직선함수 관계

- 목표변수(사망률지수)와 상관관계(직선) 높은(유의한 변수) 예측변수 사전 진단 - 산점도 행렬에서 추정 회귀식에 관측치들이 가까이 분포되어 있다면 직선 관계가 높음 - 초록색 원, 강수량(+), 교육기간(-), 비백인비율(+), 가구원수(+), 가구소득(-), SO2Pot(+)
- 상관관계 높은 예측변수 : 다중공선성 문제 발생 가능성 높음 - 비백인비율은 대부분의 예측변수와 상관관계 높음(노랑색 원)

정규분포를 따르지 않는 데이터

- 우로 치우침 : 1월 기온 우로 치우침, 교육수준, 비백인비율, 가구소득, 환경요인 변수들
- 좌로치우침 : 7월 기온, 가구원 수
- 작은 봉우리가 생기는 것은 관련 변수를 집단으로 구별하는 범주형 변수가 존재한다는 것임 - 가구소득의 경우 우측에서 작은 봉우리가 생김, 가구원수는 좌측에 작은 봉우리 졵 - 현재는 이를 판별해 볼 변수는 지역 예측변수 밖에 없음

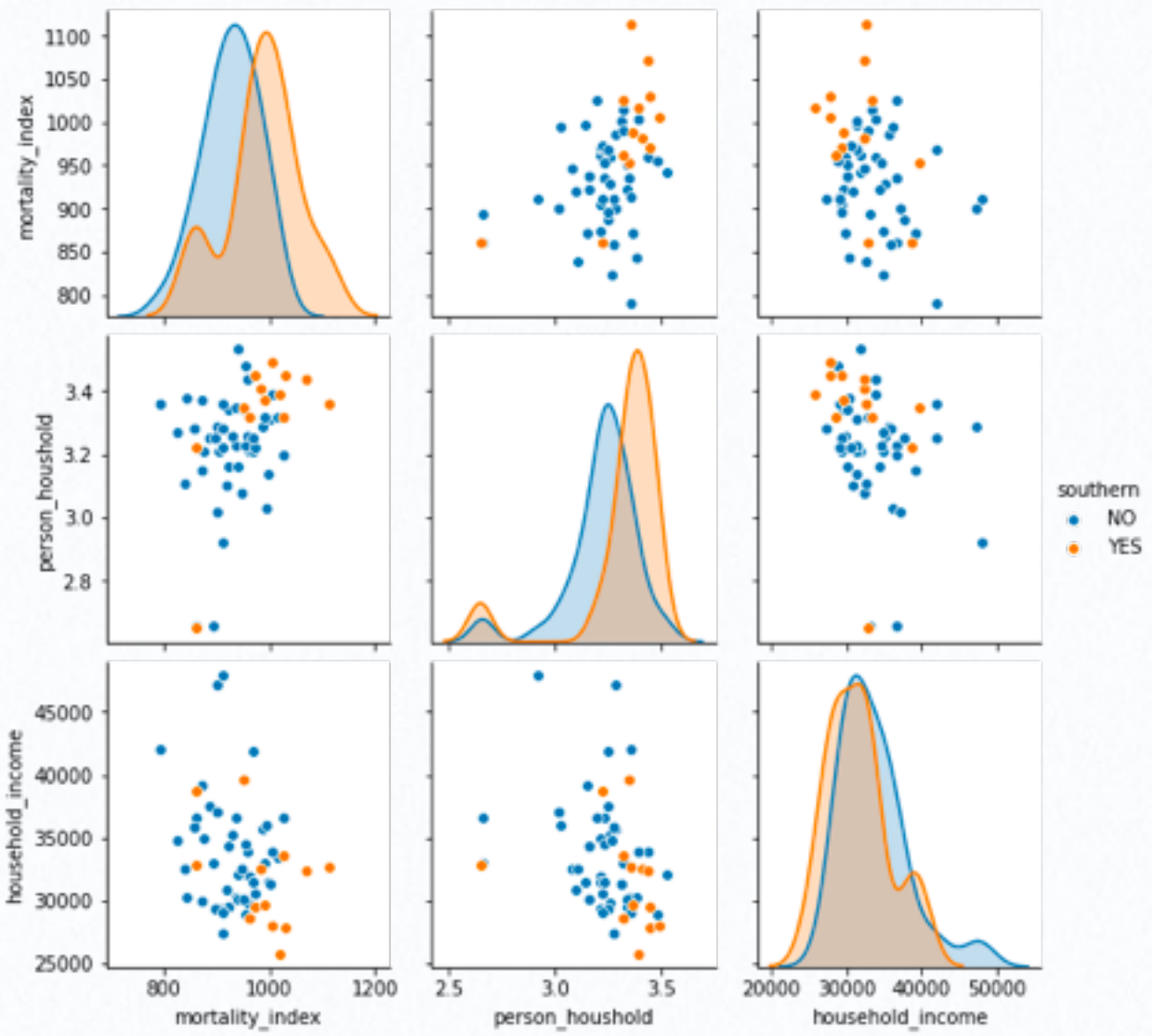


봉우리 개수(작은 봉우리라도) 2개 이상인 경우

```
import seaborn as sns
import matplotlib.pyplot as plt
smsa0=smsa[['mortality_index','person_houshold','household_income','southern']]
sns.pairplot(smsa0, hue='southern'); plt.show()
```

사망률 지수, 가구원 수 : 남부지역이 높음

가구소득은 지역별 차이를 보이지 않음



Pre-Process

체크

목표변수는 정량적 quantitative 변수이어야 한다.

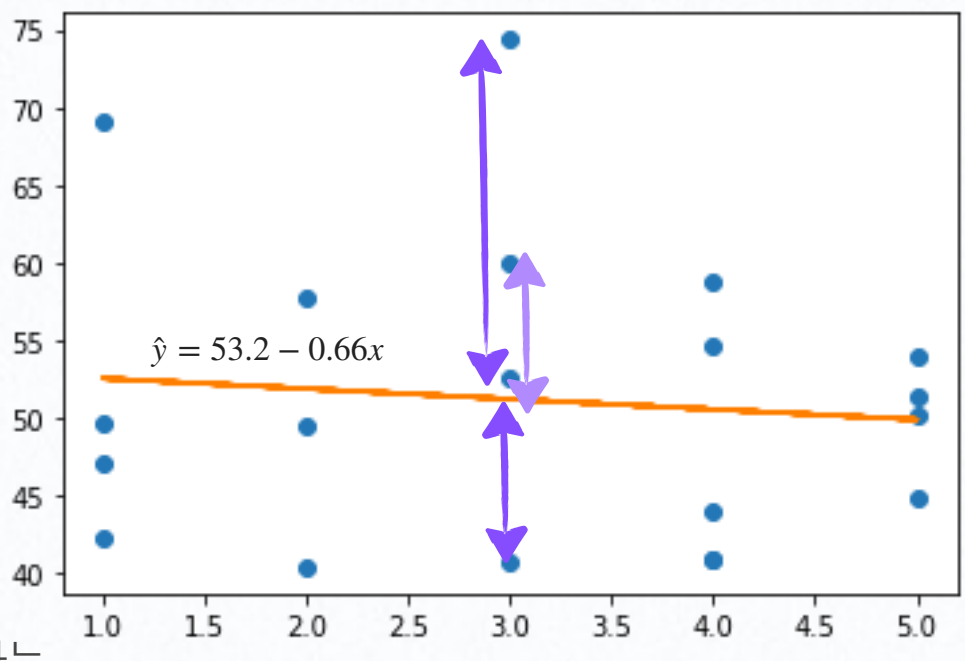
이진형, 순서형이면 로지스틱 회귀분석, 로그선형(예측변수는 모두 범주형) : 일반화 선형모형

예측변수의 종류

- 예측변수들 모두 정량적 변수 - 회귀분석 모형
- 예측변수들 모두 정성적 qualitative 변수 - 분산분석 모형
- 예측변수들이 혼합형 - (회귀분석 측면) 지시변수가 있는 회귀모형 (분산분석 측면) 공분산분석

순서형 변수

리커트 척도, 소득수준 (상, 중상, 중하, 하), 알파벳 학점 등 순서형 척도인 경우 숫자형 변수처럼 사용가능하다. 그러나 (5점) 리커트 척도는 등간척도이므로 숫자화 하는데는 문제가 없으나 소득수준(상=4, 3, 2, 1=하???), 알파벳 학점 등을 숫자화 하는 규칙은 다소 자의적이다. 그러므로 이런 경우 아래에서 논의할 범주형 변수의 변환 규칙을 적용하면 된다.



등간척도의 경우 그 값을 그대로 회귀모형에 사용하는 것은 문제가 없으나 동일 척도 값에서 여러번 Y가 관측되므로 회귀식의 적합하지 않아 발생하는 변동과 반복 관측하여 발생하는 변동이 혼합되어 실제 회귀식이 적합한지의 여부가 왜곡된다. 이에 대한 분석을 작합성 결여 분석이라 한다.

Lack of Fits 적합성 검정

회귀모형 : 총변동=회귀변동+오차변동으로 나뉜다.
오차변동을 순수변동+적합결여변동으로 나눈 후 순수변동으로 회귀계수의 유의성을 검정한다.

$$\sum (y_i - \hat{y}_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + LOF$$

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5141	5141	3.14	0.110
Residual Error	9	14742	1638		
Lack of Fit	4	13594	3398	14.80	0.006
Pure Error	5	1148	230		
Total	10	19883			

회귀오차변동 14742=분산분석오차변동(순수변동) 1148 + 적합결여변동 13954로 분리된다. 자유도 동일하게 분리된다.

회귀모형 유의성 검정

귀무가설 : $y = a + bx$는 적합하지 않다. $b = 0$

대립가설 : $y = a + bx$ 적합하다. $b \neq 0$

검정통계량 : $\frac{MSR}{MSRE} = 5141/1638 = 3.14$ 대신 $\frac{MSLF}{MSPE} = 3398/230 = 14.8$을 사용한다.

그러므로 일반 회귀분석 결과는 예측변수 X는 목표변수 Y를 유의적으로 설명하지 못한다고 하지만 적합성 결여 검정을 하면 유의한 (직선) 관계가 있다고 한다.

Analysis of Variance

Source	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\hat{y}_{ij} - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Residual Error	$n - 2$	$SSE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$	$MSE = \frac{SSE}{n-2}$	
Lack of Fit	$c - 2$	$SSLF = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_{ij})^2$	$MSLF = \frac{SSLF}{c-2}$	$F^* = \frac{MSLF}{MSPE}$
Pure Error	$n - c$	$SSPE = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$MSPE = \frac{SSPE}{n-c}$	
Total	$n - 1$	$SSTO = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$		

$E(MSPE) = \hat{\sigma}^2$, $E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$

$$E(MSLF) = \sigma^2 + \frac{\sum n_i(\mu_i - (\beta_0 + \beta_1 X_i))^2}{c-2}$$

범주형 변수 클래스

선형모형는 수학함수이므로 X는 숫자형(정량적)변수이어야 Y를 숫자로 출력할 수 있다. 예를 들어 $y = 3.2 + 0.2x$ 일차 함수에서 Y가 학점이고 X가 성별이라고 하자. X가 남자인 경우 어떤 값을 넣어야 하나? $3.2 + 0.2 * Male = ?$

이처럼 예측변수가 숫자가 아닌 범주인 경우 범주 그대로를 모형에 넣는 것은 의미가 없다 그러므로 범주를 숫자로 변환하여 넣어야 한다. 어떻게 변환할 것인가? 남자=1, 여자=2, 그러면 남자인 경우 Y=3.4 여자는 3.6이다. 그럼 남자=0, 여자=1을 사용하면? 추정식은 $y = 3.4 + 0.2x$가 될 것이다.

이처럼 범주형 변수는 숫자형으로 변환 하여 사용해야 하며 변환 규칙은 일대일 규칙이면 충분하고 어떤 값으로 변환할지는 분석자의 결정이다. 가장 널리 사용하는 것은 사용이 편리한 (0,1) 규칙이다. 이를 더미 변수라고 한다.

성별은 (0,1)로 하면 되지만 소득수준 4개 범주(상, 중상, 중하, 하)에 대한 변환규칙? 범주의 수준 개수보다 1개 적은 더미 변수가 필요하다. 성별의 경우 (남, 여) 2개이므로 1개 더미변수로 구별이 가능하였다. 4개 수준을 가진 소득수준 범주는 3개의 더미가 필요하다. 더미1=1, 더미2=0, 더미3=0 (상), 더미1=0, 더미2=1, 더미3=0 (중상), 더미1=0, 더미2=0, 더미3=1 (중하), 더미1=0, 더미2=0, 더미3=0 (하)로 구별하면 된다.

교사연봉(salary) 영향을 미치는 예측변수로 측정형인 경력(X_1)과 범주형인 대학원 졸업여부 ($X_2 = 1(yes)$,$0(no)$)를 설정하였다.

$$S = b_0 + b_1X_1 + b_2X_2 + b_3(X_2X_3) + e$$

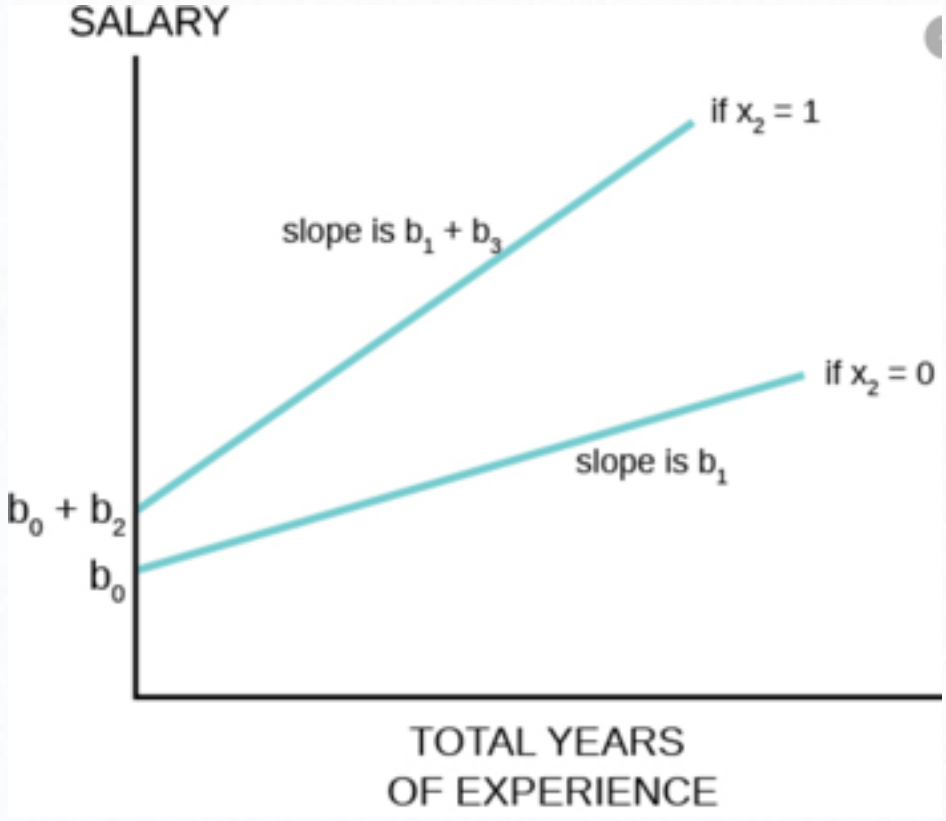
회귀계수 b_2는 대학원 졸업여부(X_1)에 따라 절편의 증감, b_3는 대학원 졸업여부(X_1)에 따라 기울기의 변동이 발생한다.

즉 $X_1 = 0$(대학원 미졸업자인 경우)

$$S = b_0 + b_2X_1$$

즉 $X_1 = 1$(대학원 졸업자인 경우)

$S = (b_0 + b_2) + (b_1 + b_3)X_1$가 된다.

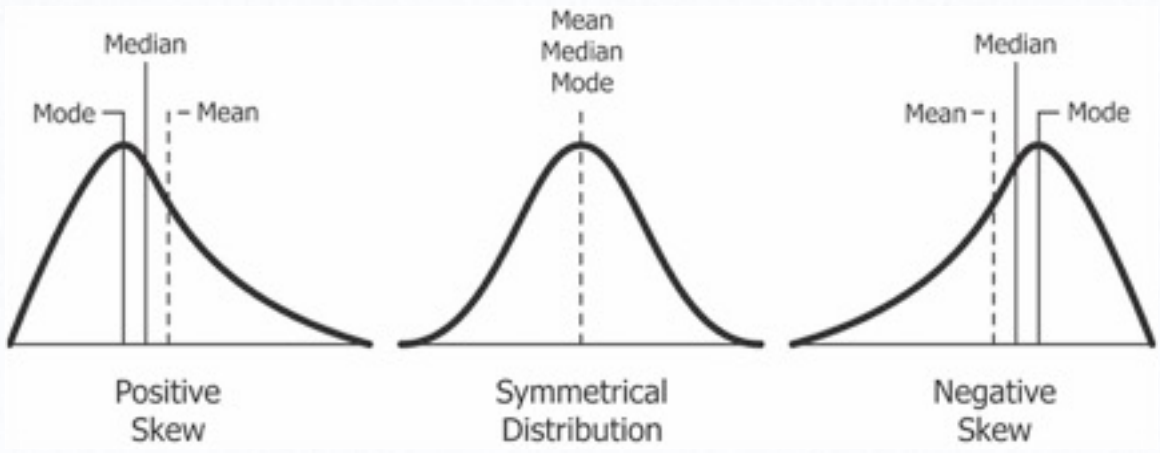


대학원 졸업여부에 의한 경력의 증가가 연봉에 미치는 기울기 차이는 b_3의 유의성에 의해 검정된다.

변수 정규성 검정

치우침 skewness

오른쪽 꼬리가 길거나 왼쪽 꼬리가 길게 되면 데이터의 실증적 규칙이 성립하지 않음



- 실증적 규칙 : 평균 ±2(표준편차) 구간에 데이터의 95%가 내재되어 있음
- 그러나 치우친 경우에는 Cheychev 부등식에 의해 평균 ±2(표준편차) 구간에는 적어도 75% 포함되어 있음 - 즉 데이터의 분산이 커지게 된다.

통계 추론은 데이터의 정규성을 가정하는 경우가 대부분이다.

통계량 활용 - 치우침 판단

- 수리 왜도 skewness : $\frac{E(X-\mu)^3}{\sigma^2}$
- EDA 왜도 : $\frac{(Q_3+Q_1-2Median)}{(Q_3-Q_1)}$
- Pearson Median 왜도 : $PS=\frac{(Mean-Mode)}{\sigma}$(제1 공식), $PS=\frac{3(Mean-Median)}{\sigma}$(제2 공식)
- 정규분포=0, 우로 치우침 +, 좌로 치우침 : 이는 통계량의 분포를 모르므로 정규분포 가설을 검정할 수 없어 시각적 판단 수준임

치우침 검정 = 정규성 검정

데이터의 분포가 이론적 정규분포를 따르는지 검정하는 적합성 검정임

- 귀무가설 : 데이터 모집단 분포는 정규분포이다
- 대립가설 : 정규분포를 따르지는 않는다 <=> 그러나 어떤 분포인지는 모른다.

(1) Shapiro Wilk W-통계량

$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$, 상수 a_i는 분산행렬을 이용하여 구함

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \le n \le 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } 12 \le n \le 2000 \end{cases}$$

(2) Kolmogorov D-통계량

- $D = \max_x \left| F_n(x) - \Phi(x) \right|$
- $\Phi(x)$ 누적정규분포, $Fn(x)$ 데이터누적분포함수

(3) Anderson-Darling AD 통계량

- $A^2 = n \int (Fn(x) - \Phi(x))^2 \left| \Phi(x)\Phi(1-x) \right|^{-1} d\Phi(x)$

어떤 방법이 적절한가? 분석자의 결정이나 Shapiro-Wilks 방법이 가장 범용적이다.

치우침 문제 해결 = 정규변환이 필요한 이유

변수가 하나인 경우

통계학의 모수적 검정방법(통계량의 샘플링분포)에는 데이터(변수) 모집단 분포는 정규분포(좌우 대칭인 분포)를 가정한다.

물론 평균, 비율에 관한 검정(차이 포함)은 대표본 이론(중심극한정리)에 의해 데이터 모집단 분포에 관계없이 표본평균과 표본비율의 샘플링 분포는 정규분포에 근사하므로 데이터 변환의 필요 없음

그러나 분산에 대한 추론에서는 표본분산의 샘플링분포는 카이제곱분포를 이용한다. 그러려면 모집단 분포는 정규분포를 따라야 한다

- 모집단 분포 population dist.: 데이터가 추출된 모집단의 분포 $x \sim f(x)$, 일반적으로 알 수 없으며 필요한 경우 분포 가정을 함,모집단 분포에 대한 관심보다는 모집단 분포의 특성 값 (모수, parameter)을 추론(가설검정, 추정) 하게 됨

- 표본 분포 sample dist. : 확률표본 데이터 분포 $f(x)$, 모집단의 분포와 동일함 - 히스토그램을 그려 분포함수의 형태를 볼 수 있음, 이론적 분포(가정한 분포)와 같은지 검정은 분포 적합성(goodness od fits) 검정 실시
- 샘플링 분포 : 표본 데이터로부터 계산된 통계량의 분포를 의미한다. 통계량이 추정에 사용되면 추정량, 가설 검정에 사용되면 검정통계량이라 한다. (예) 표본평균은 모평균의 추정량이며, 표본평균의 함수인 $(\bar{x} - \mu_0)/(s/\sqrt{n})$ 은 모평균 가설검정 통계량임. 통계량의 분포를 샘플링분포라고 하고 모집단의 분포와 관계없이 표본평균의 샘플링분포는 (중심극한정리에 의해)정규분포에 근사한다.
- 표본이 대표본이 아닌 경우(모집단 분포의 치우침의 크기에 좌우하나 표본크기 20~30이면 대표본) 표본평균의 샘플링 분포는 모집단이 정규분포(적어도 좌우 대칭)이어야 표본평균이 정규분포를 따르고 t-검정방법을 사용한다.
- 소표본이고 모집단이 정규분포를 따른다는 가정이 없다면 모수적 검정방법(통계량 샘플링분포(t-분포)를 이용한 추론)을 사용할 수 없고 비모수적방법(non-parametric, dist. free 분포자유)을 사용해야 한다.

모수적 방법을 사용하려면 모집단의 분포가 정규분포를 따라야 하므로 (1) 모집단의 분포가 정규분포를 따르는지 검정? <=> 표본 데이터의 정규성 검정 (2) 정규성 검정을 만족하지 않으면 변수변환(변수 분포의 정규분포 검정)

선형모형

- $y = f(x) + e$ 통계모형에서 오차항의 분포는 정규분포를 가정한다.
- 오차항의 분포가 정규분포를 따라야 모형의 모수에 대한 모수적 방법(통계량의 샘플링 분포, 회귀계수에 대한 t-검정, 분산분석 F-검정) 가능하다
- 통계모형에서는 목표변수 y만 확률변수이고 예측변수(input)는 결정변수(확률분포함수를 가지지 않음)이므로 오차항이 정규분포이면 목표변수도 정규분포에 근사해야 한다

그러나 통계 선형모형에서 모든 변수(X, Y)는 정규분포에 근사해야 모형의 적합성이 높아짐 - 그리하여 치우침이 있는 데이터의 경우 미리 변환하여 모형에 삽입하게 된다. (예) 소득, 가격, 수능점수

수리통계

- 확률변수 X의 확률밀도함수가 $f(x)$를 따른다. 그렇다면 X의 함수 $g(X)$가 갖는 확률밀도함수는?
- $U \sim U(0, 1)$ 균일분포이면, $-\frac{1}{\lambda}\ln(1 - U) \sim \exp(\lambda)$ 지수분포를 따른다.

- 만약 $Z \sim SN(0,1)$표준정규분포, $Z^2 \sim \chi^2(1)$ 카이제곱 분포를 따른다

간단한 정규변환 normal transformation 방식

- 우로 치우침 : $\sqrt{x} \rightarrow x^{1/3} \rightarrow ln(x) -> log(x)$
- 좌로 치우침 : $x^2 \rightarrow x^3$

Tukey Ladder of Power $x' = x^\lambda$

- 좌로 치우침 : $\lambda = 2(x^2) < 3(x^3)$
- 우로 치우침 : $\lambda = 1/2(\sqrt{n}), 1/3(x^{1/3}), 0(ln(x)) -> -1/2(-1/\sqrt{x}), -1(-1/x), -2(-1/x^2)$

단일변수 정규변환

- λ 값을 변환하면서 데이터의 정규성 검정을 실시하여 최적의 λ 값을 찾음

ANOVA (선형모형)

- 잔차의 정규성 검정 -> 최적의 λ 값을 찾고 목표변수를 변환하여 다시 분석함

선형모형

- 목표변수만의 변환은 예측변수와의 함수관계 왜곡이 되므로 분석 전에 모든 예측변수 포함 모든 변수를 개별적 정규변환이 필요함

Box-Cox tranformation $x' = \frac{(x^\lambda - 1)}{\lambda}, x > 0 \mid x' = ln(x), x = 0$

- Tukey 변환가 동일하지만 최적의 λ 값은 MLE 방법에 의해 찾음
- 단일변수 정규성 변환에서는 Tukey 방법이 ANOVA에서는 Box-Cox 방법이 적절함

Lambda (λ)	Transformed Distribution (Y')
-2	$Y' = \frac{1}{Y^2}$
-1	$Y' = \frac{1}{Y^1}$
-0.5	$Y' = \frac{1}{sqrt(Y)}$
0	$Y' = log(Y)$
0.5	$Y^1 = sqrt(Y)$
1	$Y' = Y$
2	$Y' = Y^2$

파이썬 프로그램[정규성 검정]

- 귀무가설 : 데이터는 정규분포를 따른다.
- 대립가설 : 데이터는 정규분포를 따르지 않는다.

```
from scipy.stats import shapiro
shapiro(smsa.jan_temp)
```

출력결과는 검정통계량과 유의확률이다. (0.9217209815979004, 0.0010030175326392055)

유의확률이 0.05보다 작으므로 귀무가설은 기각되어 1월 기온은 정규분포를 따르지 않는다.

SMSA 정규성 검정

```
from scipy.stats import shapiro
for k in range(0,15):
  if shapiro(smsa.iloc[:,k])[1]<0.05:
    print(k,'번째',smsa.columns[k],'Not Normal')
```

```
0 번째 jan_temp Not Normal
2 번째 humidity Not Normal
3 번째 rainfall Not Normal
6 번째 pop_density Not Normal
7 번째 non_white_ratio Not Normal
9 번째 population Not Normal
10 번째 person_houshold Not Normal
11 번째 household_income Not Normal
12 번째 HCPot Not Normal
13 번째 NOxPot Not Normal
14 번째 SO2Pot Not Normal
```

SMSA 측정형 데이터 15개의 정규성 검정 결과 정규분포를 따르지 않는 변수 리스트이다.

정규변환 Normal Transformation

```
from scipy import stats
xt, lmbda = stats.boxcox(smsa.jan_temp)
```

정규분포를 따르지 않는 데이터는 정규변환을 하는 것이 적절하다.

Box-cox 정규변환 결과 변환된 데이터와 람다 값을 출력한다.

```
1  print(lmbda)
```

```
0.07425875013260441
```

```
[85] 1  print(xt)
```

```
[3.73417559 3.53058515 3.82569238 4.39918394
 3.8692799  3.8692799  3.58438789 3.73417559
 4.03115702 3.78069058 3.91154259 4.42836668
 3.73417559 3.58438789 3.58438789 4.24360516
 3.82569238 3.91154259 3.95256155 4.61759172
 4.93512314 3.35509312 2.72894876 4.24360516
 3.99241058 3.95256155 3.82569238 4.17627638
 4.21034036 3.63615398 3.95256155 4.66740279
 4.24360516 3.78069058 3.58438789 3.68603753
 3.95256155 3.99241058 3.58438789 3.99241058
```

```
1  (smsa.jan_temp**(lmbda)-1)/lmbda
```

```
0    3.734176
1    3.530585
2    3.825692
3    4.399184
```

한번에 모든 비정규 데이터 정규변환 및 저장

```
from scipy import stats
from scipy.stats import shapiro
for k in range(0,15):
  if shapiro(smsa[smsa.columns[k]])[1]<0.05:
  smsa[smsa.columns[k]]=stats.boxcox(smsa[smsa.columns[k]])[0]
```

```
[212] 1  smsa_nor=smsa.set_index(['state_nm','city_nm'])
      2  smsa_nor.head()
```

		jan_temp	july_temp	humidity	rainfall	mortality_index	education	pop_density
state_nm	city_nm							
OH	Akron	3.734176	71	43.268170	55.674897	921.87	11.4	10.042724
NY	Albany-Schenectady-Troy	3.530585	72	41.903399	53.842582	997.87	11.0	10.468087
PA-NJ	Allentown	3.825692	74	39.847721	70.629792	962.35	9.8	10.460501

필드에서 정규변환 방법

우로 치우침만 제곱근 혹은 로그 변환으로 정규화 하자.

로그변환, 제곱근 변환 중 유의확률이 높은 변환(정규분포에 적절함)하고 변환하여 원 데이터에 _log, _sqrt로 저장하였음, 유의확률이 0.05넘지 않으면

```
import numpy as np
def rskewed(k):
  if (shapiro(np.sqrt(smsa.iloc[:,k]))[1]<0.05) & (shapiro(np.log(smsa.iloc[:,k]))[1]<0.05):
        print('No Appropriate Normal Transformation (right skewed)')
  else:
    if(shapiro(np.sqrt(smsa.iloc[:,k]))[1]<shapiro(np.log(smsa.iloc[:,k]))[1]):
      print(k,':',smsa.columns[k], '로그변환 : 유의확률',shapiro(np.log(smsa.iloc[:,k]))[1])
      smsa[smsa.columns[k]+str('_log')]=np.log(smsa.iloc[:,k])
    else:
      print(k,':',smsa.columns[k], '제곱근변환 : 유의확률',shapiro(np.sqrt(smsa.iloc[:,k]))[1])
      smsa[smsa.columns[k]+str('_sqrt')]=np.sqrt(smsa.iloc[:,k])
```

좌로 치우침은 해결 제곱이 가장 적절한데 심하지 않는 경우에는 하지 않아도 된다.

봉우리가 있는 경우 정규변환이 해결되지 않는다.

우로 치우침 문제 해결

```
for k in [0,6,7,9,11,12,13,14]:
  rskewed(k)
```

```
No Appropriate Normal Transformation(right skewed)
6 : pop_density 로그변환 : 유의확률 0.6019314527511597
7 : non_white_ratio 제곱근변환 : 유의확률 0.24234969913959503
9 : population 로그변환 : 유의확률 0.388794809579849 24
No Appropriate Normal Transformation(right skewed)
No Appropriate Normal Transformation(right skewed)
13 : NOxPot 로그변환 : 유의확률 0.45190250873565674
No Appropriate Normal Transformation(right skewed)
```

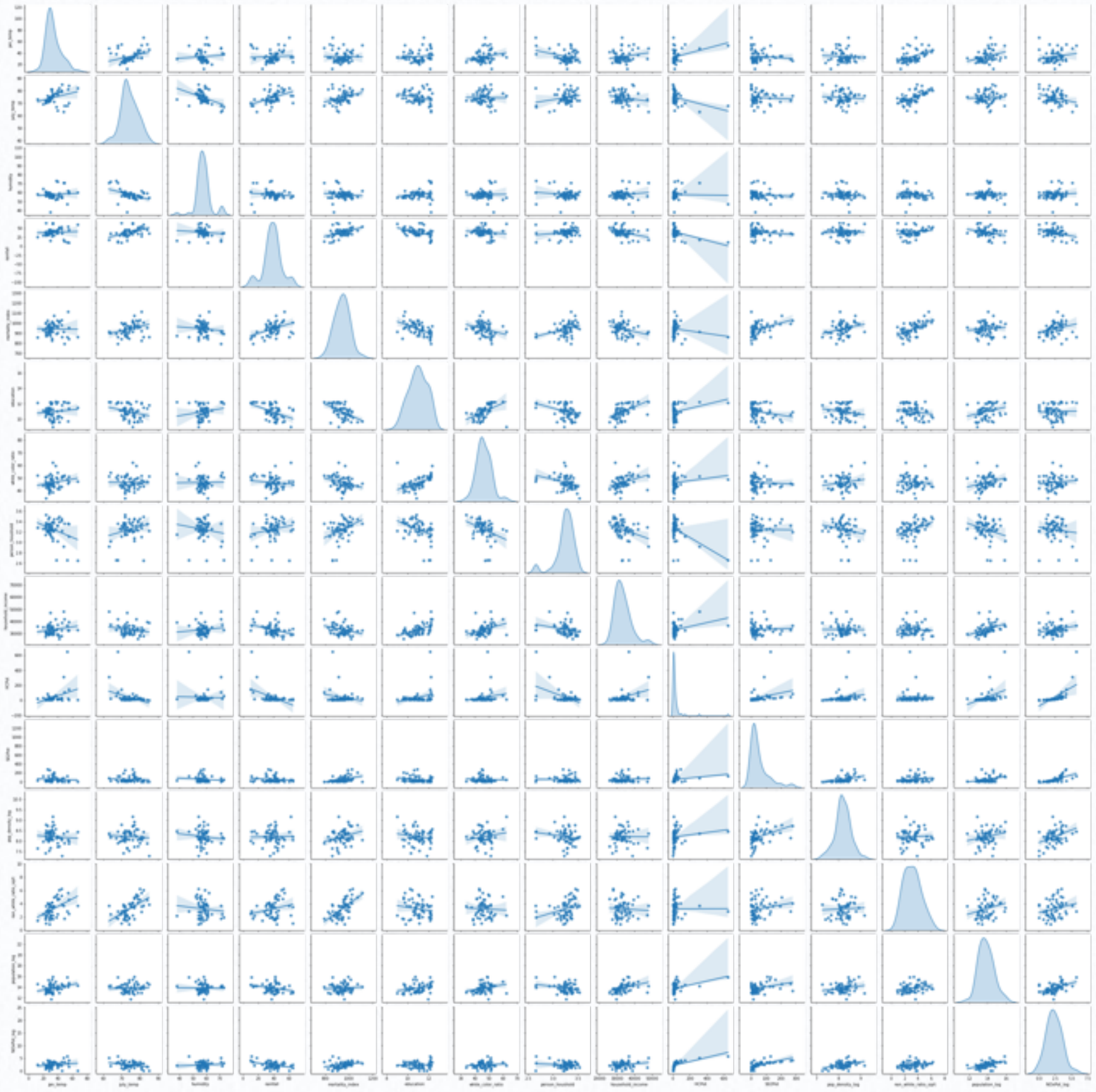
향후 선형(예측모형) 분석은 전처리 완료된 데이터 smsa_clean 활용합니다.

```
smsa_clean=smsa.drop(['pop_density','non_white_ratio','population','NOxPot'],axis=1)
smsa_clean.set_index(['state_nm','city_nm'],inplace=True)
smsa_clean.columns
```

```
Index(['jan_temp', 'july_temp', 'humidity', 'rainfall', 'mortality_index',
       'education', 'white_color_ratio', 'person_houshold', 'household_income',
       'HCPot', 'S02Pot', 'southern', 'city_nm', 'state_nm', 'pop_density_log',
       'non_white_ratio_sqrt', 'population_log', 'NOxPot_log'],
      dtype='object')
```

정규변환 후 최종 선형분석 데이터(usa_smsa) 산점도 행렬

```
6 : pop_density 로그변환 : 유의확률 0.6019314527511597
7 : non_white_ratio 제곱근변환 : 유의확률 0.24234969913959503
9 : population 로그변환 : 유의확률 0.38879480957984924
13 : NOxPot 로그변환 : 유의확률 0.45190250873565674
```



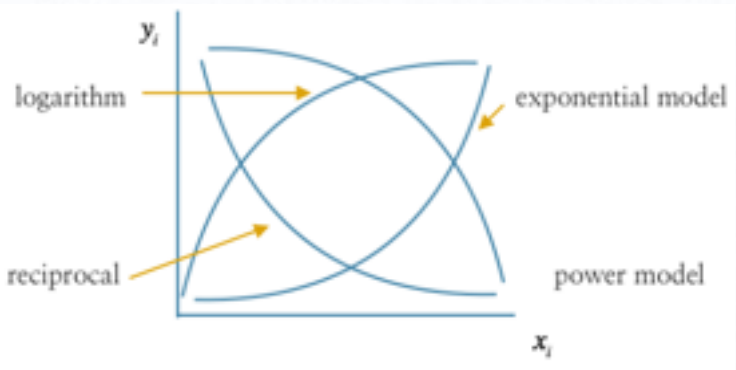
함수의 선형성 검토

개념

- 목표변수와 예측변수가 선형이 아닌 경우 - 선형회귀모형이므로 문제 해결 필요

진단도구

- 산점도 행렬 : 목표변수와 예측변수 간 개별 산점도 - 직선관계 발견



해결책 : 선형 변환 Linear Transformation

모형	식	Y변환	X변환
Power	$y = ax^b$	$ln(y)$	$ln(x)$
Exponetial	$y = ae^{bx}$	$ln(y)$	x
logarithmic	$y = ln(ax^b)$	y	$ln(x)$
reciprical	$y = \frac{1}{a+bx}$	$\frac{1}{y}$	x
Square Root	$y = a + b\sqrt{x}$	y	$\sqrt{x}$
Square	$y = a + bx^2$	y	x^2

한계

단순회귀모형에서는 목표변수의 선형변환이 가능하나, 다중회귀모형에서는 불가능하다. 왜냐하면 한 예측변수와의 선형성을 수정하면 다른 예측변수와의 선형성 관계가 왜곡될 수 있기 때문이다.

다중모형에서는 선형성 변환은 예측변수에 국한되며, 게다가 빅데이터, 예측변수가 많은 경우 선형성 진단은 필요하지 않다.

Correlation Aanlysis

개념

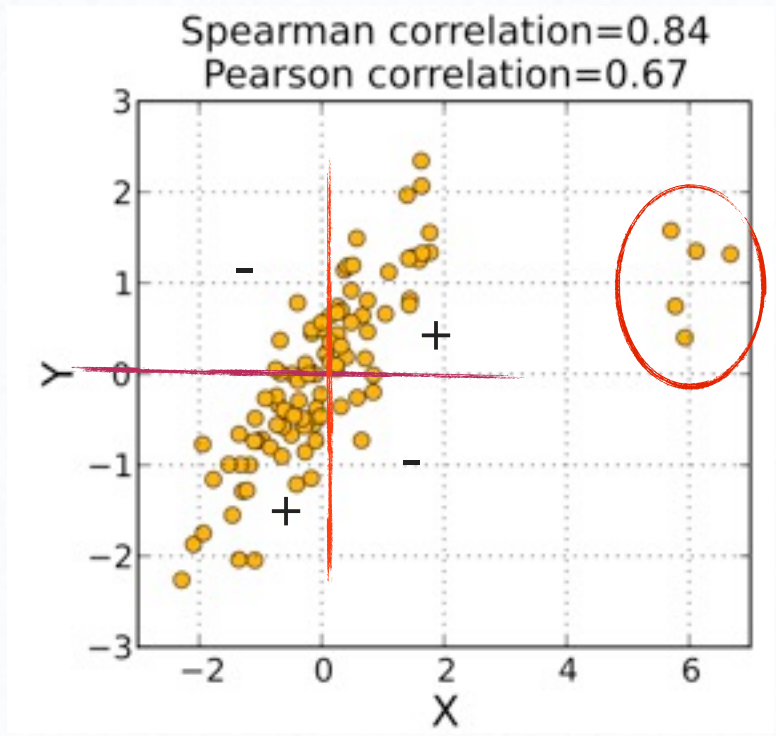
두 양적(순서형 포함) 변수(X, Y)간의 직선 관계 정도를 계수로 측정함

직선관계가 유의하다는(한 변수가 증가하면 다른 변수도 직선적으로 증가하거나 감소함) 것은 두 변수가 유사하다는 의미 - 변수의 유사성 척도

상관계수

Karl Peason 공식

- $\rho = \frac{COV(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$: 모집단
- $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
- 분모는 확률변수의 표준편차이므로 상관계수의 부호를 결정하는 분자항이다. $(x_i - \bar{x})(y_i - \bar{y})$의 부호는 아래 그림(수평선은 Y의 평균, 수직선은 X의 평균, 오른쪽의 관측치 5개를 제외한 경우)에서 시각적으로 확인할 수 있음.



Spearman 순위 상관계수

- (방법 1) $r_s = Corr(R_{X_i}, R_{Y_i})$ where R_{X_i}는 X_i의 순위이며, R_{Y_i}는 Y_i의 순위이다.
- (방법 2) $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \ d_i = R_{X_i} - R_{Y_i}$

- 대부분의 데이터 범위 밖에 있는 관측치(타원형 내 관측치)는 상관계수 값을 높이는 역할을 한다. 그러므로 상관계수를 계산하기 전에 반드시 산점도를 그려 데이터의 범위를 많이 벗어난 관측치가 있는지 확인하여 상관분석의 활용도를 높일 필요가 있음.

Kendall τ 상관계수

- $\tau = \frac{\#of\_concordant\_pairs - \#of\_disconcordant\_pairs}{n(n-1)/2}$
- concordant = 만약 $(x_i > x_j)$, $(y_i > y_j)$이거나 $(x_i < x_j)$이면, $(y_i < y_j)$이면 두 관측치는 concordant 쌍이라 함
- τ 값이 클수록 데이터 순위의 일치도는 높아지므로 상관관계가 높아진다.

상관계수 유의성 검정

가설

- 귀무가설 : 두 변수의 **직선** 상관관계는 유의하지 않다. <=> 서로 독립이다. $\rho = 0$
- 대립가설 : 두 변수의 **직선** 상관관계는 유의하다. $\rho \neq 0$

데이터 검증

- 데이터는 이변량 정규분포에 근사해야 한다. 단 n>20 인 대표본에서는 문제 없음
- 산점도를 그려 데이터 범위(X-) 밖의 관측치 존재 여부를 체크한다. - 존재한다면 제외하거나 활용 시 주의해야 한다.

검정통계량

- $TS = \frac{r}{\sqrt{(1-r^2)(n-2)}} \sim t(n-2)$, n=표본크기, r=상관계수

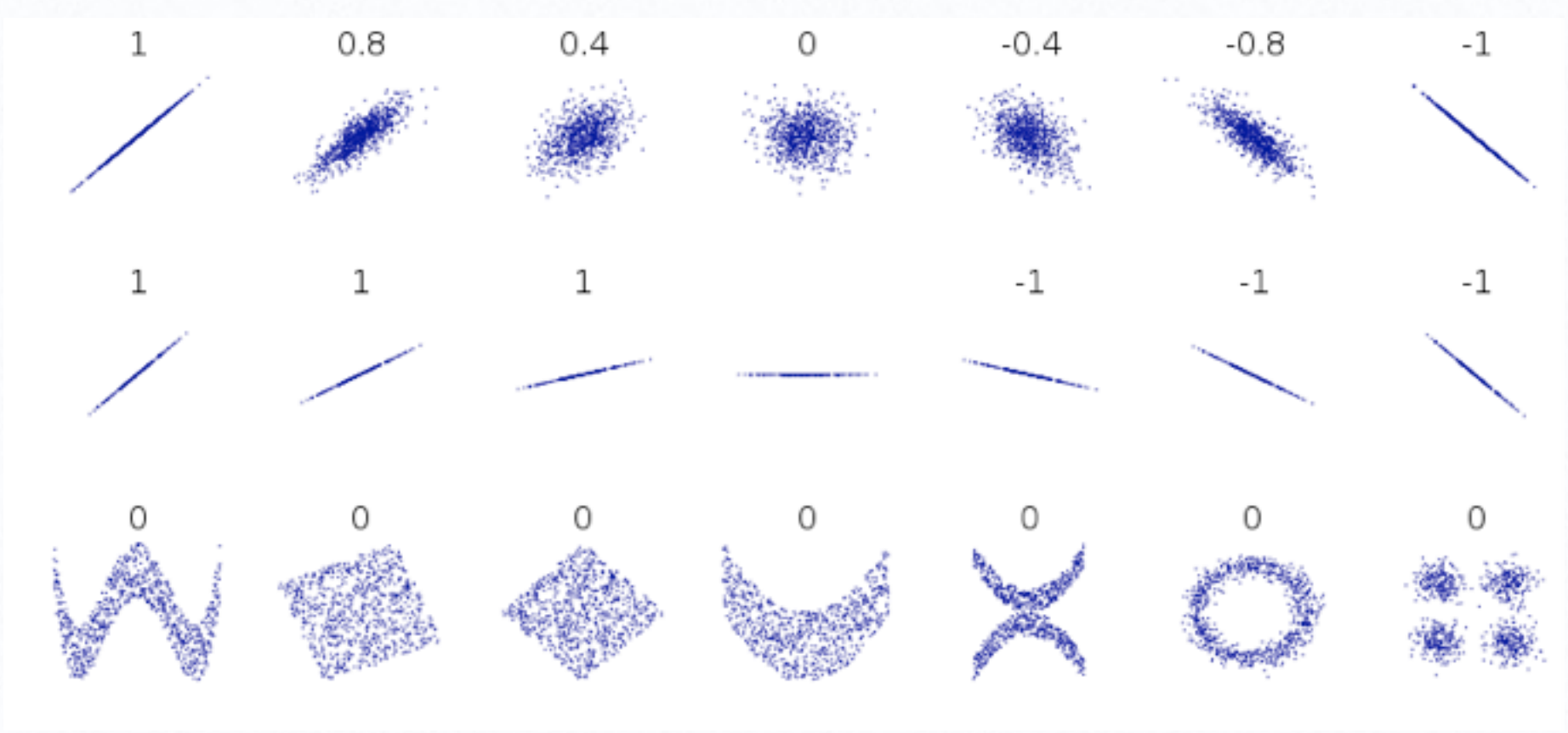
(만약) 귀무가설이 $\rho = \rho_0 \neq 0$ (임의의 집단 상관계수와 동일한지 검정한다면)이라고 하면

- (예) 미국 대학생 성적과 공부시간의 상관계수는 0.7($\rho_0 = 0.7$)이다. 한국 대학생?
- 검정 통계량 : $TS = \frac{1}{2}\ln(\frac{1+r}{1-r}) \sim Normal(\frac{1}{2}\ln(\frac{1+\rho}{1-\rho}), \frac{1}{n-3})$

결론

- 유의확률 $P(t(n-2) > \left|TS\right|)$이 유의수준보다 작다면 귀무가설을 기각하여 상관관계의 유의하다고 결론내리고 표본상관계수의 부호를 이용하여 해석
- 귀무가설이 기각, 표본상관계수 부호 + => 두 변수는 양의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 증가(감소)한다.
- 귀무가설이 기각, 표본상관계수 부호 - => 두 변수는 음의 상관관계가 있고 한 변수의 값이 증가(감소)하면 다른 변수의 값도 감소(증가)한다.

상관계수 해석



*) 출처 : 위키피디아

- –1과 1사이의 값이다.

- 1에 가까우면 양의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값도 증가(감소)한다.
- –1에 가까우면 음의 선형 상관 관계가 존재한다. 한 변수의 값이 증가(감소)하면 다른 변수 값은 감소(증가)한다.
- 두 변수의 상관 관계가 높다는 것은 두 변수가 동일한(comparable) 개념을 측정한다는 의미도 담고 있다(두 변수가 유사함). 그러므로 변수를 축약하거나 개체를 분류하는데 사용되는 다변량 분석에서는 공분산, 혹은 상관계수 개념 사용

Rule of Thumb for Interpreting the Size of a Correlation Coefficient

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

단순 회귀분석과 관계 : $Y = a + bX + e$,

- 독립변수 X가 Y에 선형적 영향을 미치는지 검정 <=> 기울기 b=0(영향을 미치는 않음) 유의성 검정 <=> 상관계수의 유의성 검정과 동일
- 회귀계수 b의 부호와 상관계수 r의 부호는 동일하고 $\hat{b} = \sqrt{\frac{S_{XY}}{S_{XX}}} r$ ($S_{XX} = \sum (x_i - \bar{x})^2$, $S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$)
- 모형의 적합성을 나타내는 결정계수는 상관계수의 제곱과 같다. $R^2 = \frac{SSR}{SST} = r^2$
- 상관 계수가 0에 가깝다는 것은 선형 상관 관계가 없다는 것이지 함수 관계가 없다는 것은 아니다. 두 변수는 이차식에 의한 ($Y_i = 100 + X_i^2 - 0.4X_i$) 생성된 데이터이나 상관계수는 0에 가깝다.

파이썬 상관계수

분석 데이터 정규변환 완료 데이터 : smsa_clean (페이지 37)

```
Index(['jan_temp', 'july_temp', 'humidity', 'rainfall', 'mortality_index',
       'education', 'white_color_ratio', 'person_houshold', 'household_income',
       'HCPot', 'S02Pot', 'southern', 'city_nm', 'state_nm', 'pop_density_log',
       'non_white_ratio_sqrt', 'population_log', 'NOxPot_log'],
      dtype='object')
```

상관계수 계산

```
smsa_cor=smsa_clean.corr()
smsa_cor.head(3)
```

	jan_temp	july_temp	humidity	rainfall	mortality_index	education
jan_temp	1.000000	0.310871	0.089363	0.060369	-0.019908	0.115632
july_temp	0.310871	1.000000	-0.441534	0.476700	0.329009	-0.277496
humidity	0.089363	-0.441534	1.000000	-0.110401	-0.110262	0.195123

목표변수와 상관계수 높은 예측변수 출력 및 subset 만들기

목표변수와 상관계수가 0.4 이상인 예측변수 선택하였음 : 1개 예측변수는 정규변환한 결과임.

```
cor_target=abs(smsa_cor['mortality_index'])  #Selecting highly correlated features
cor_target[cor_target>0.4].index.values.tolist()
```

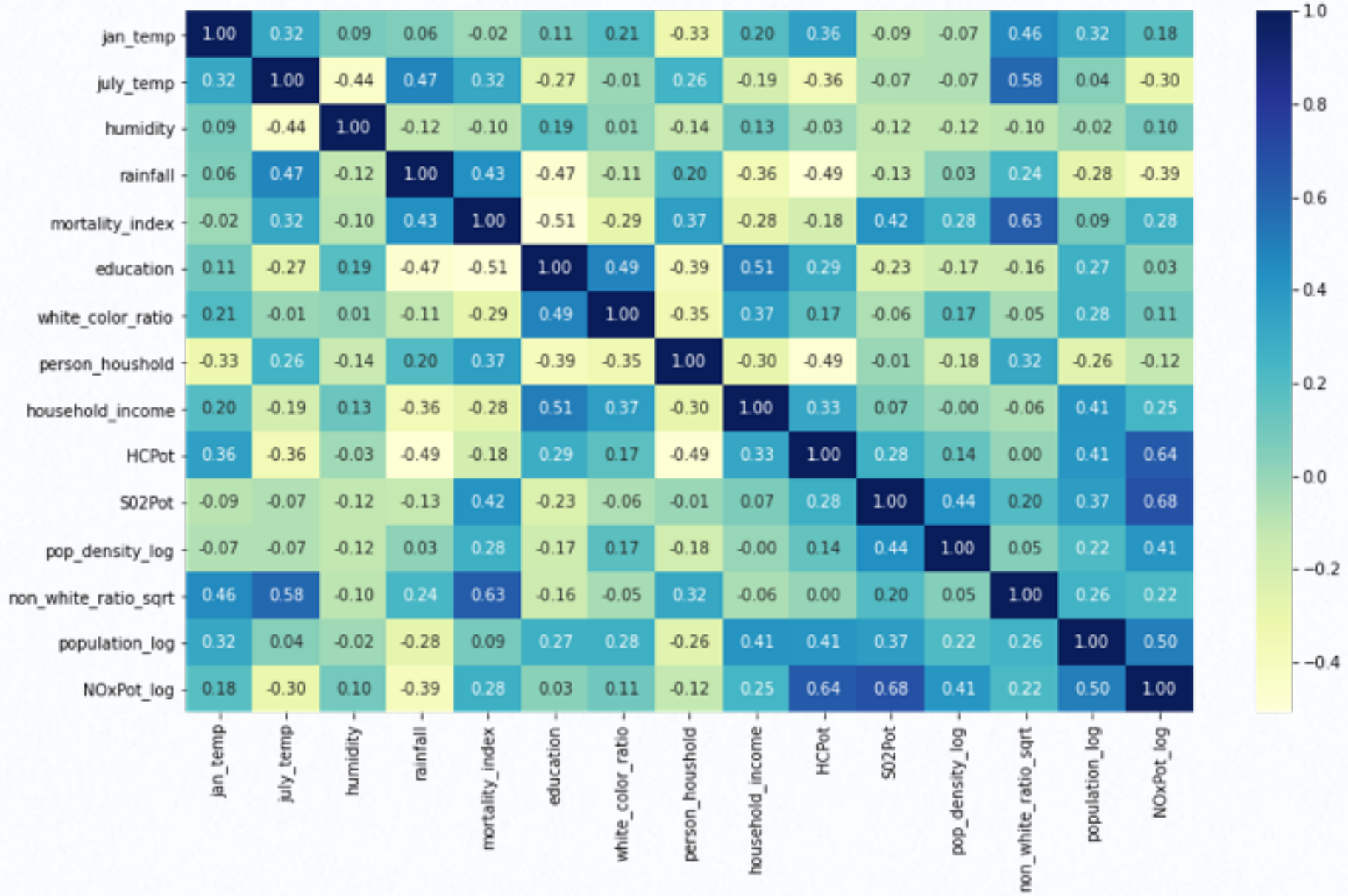
```
['rainfall', 'mortality_index', 'education', 'S02Pot', 'non_white_ratio_sqrt']
```

```
usa_smsa_sub=usa_smsa[cor_target[cor_target>0.4].index.values.tolist()]
usa_smsa_sub.columns
```

```
Index(['rainfall', 'mortality_index', 'education', 'S02Pot',
       'non_white_ratio_sqrt'],
      dtype='object')
```

상관계수 산점도 행렬 그리기

```
import seaborn as sn
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (14,8)
sn.heatmap(smsa_cor,annot=True,fmt=".2f",cmap="YlGnBu")
plt.show()
```



Estimation

Ordinay Least Squares 최소자승추정법

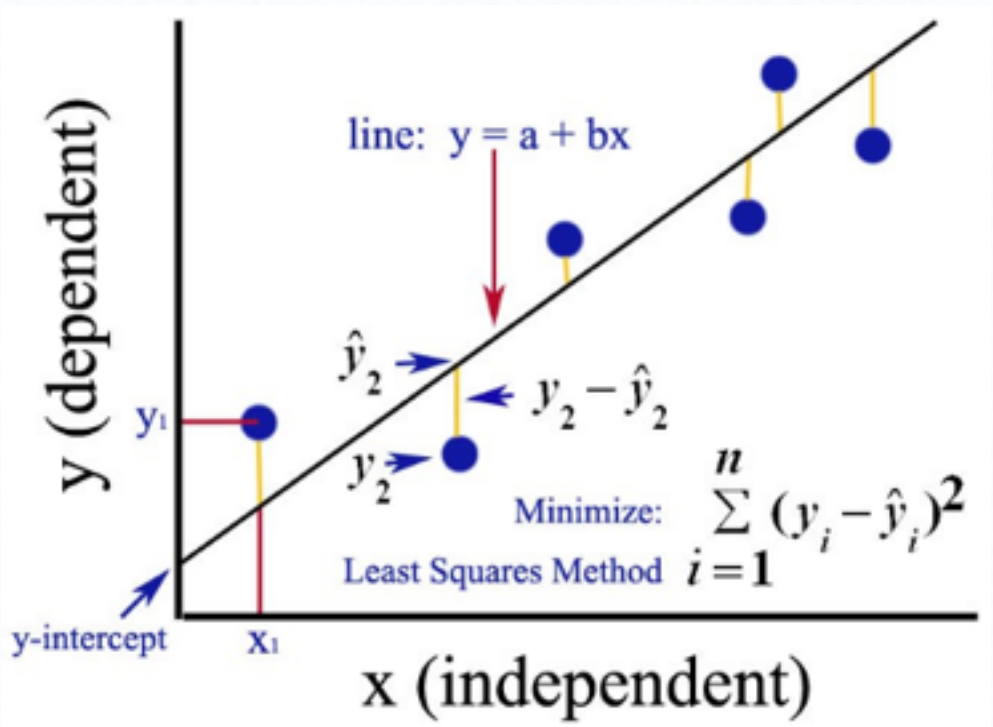
개념

Karl Pearson이 Galton의 자녀, 부모키 관계식 유도를 위하여 어버지, 성인아들 관계식 도출

방법

$$\min_{\alpha,\beta_1,\beta_2,\ldots} \sum (e_i)^2 = \sum (y_i - \hat{y}_i)^2$$

점들을 대표하는 (가장 적합하다고 판단되는) 추정식을 어떻게 구할 것인가?



수평, 수직 편차 중 왜 수직 편차인가?

- 수평편차는 예측변수의 변동에 대한 척도이고 수직편차는 목표변수의 변동을 척도이다.
- 선형모형은 목표변수의 변동(값)을 예측하는 것이 목표이므로 수직 편차를 최소화 하는 것이 점들을 대표하는 회귀식(적합식)이다.

왜 제곱인가?

- 수평 편차 $(y_i - \hat{y}_i)$의 합은 0이다. 그럼 왜 절대값을 사용하지 않나?
- 절대값은 수학적으로 다루기 쉽지 않으므로 1)회귀선에서 많이 벗어날수록 높은 페널티를 주고 2)수학적으로 다루기 쉬운 제곱을 사용한다.

점추정

행렬모형 : $\underline{y} = X\underline{b} + \underline{e}$, (가정) $\underline{e} \sim N(\underline{0}, \sigma^2 I_n)$

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}_{n\times 1}, \underline{e} = \begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix}_{n\times 1}, \underline{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \cdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times p}$$

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ik} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \cdots & \sum x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik} & \sum x_{i1}x_{ik} & \sum x_{i2}x_{ik} & \cdots & \sum x_{ik}^2 \end{bmatrix} \mathbf{X}^\mathsf{T}\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ik}y_i \end{bmatrix}$$

오차항의 가정 $\underline{e} \sim N(\underline{0}, \sigma^2 I_n)$이므로 $\underline{y} \sim N(X\underline{b}, \sigma^2 I_n)$

$$\min_{\alpha,\beta_1,\beta_2,\ldots} \sum (e_i)^2 \rightarrowtail \min_{\underline{\beta}} \underline{e}'\underline{e} = \min_{\underline{\beta}} (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$$

$$Q = (\underline{y} - X\underline{b})'(\underline{y} - X\underline{b}) = (\underline{y}' - \underline{b}'X')(\underline{y} - X\underline{b}) = \underline{y}'\underline{y} - \underline{y}'X\underline{b} - \underline{b}'X'\underline{y} + \underline{b}'X'X\underline{b}$$

Q를 최소화 하는 회귀계수(모수) $\underline{b}$를 찾는 것은 미분 값이 0인 해를 찾는 것과 동일하다.

$$\frac{\partial Q}{\partial \underline{b}} = -X'\underline{y} - X'\underline{y} + 2X'X\underline{\hat{b}} = 0 \Rightarrow X'X\underline{\hat{b}} = X'\underline{y} \Rightarrow \underline{\hat{b}} = (X'X)^{-1}(X'\underline{y})$$

점추정 평균과 (추정)분산

- 불편추정량 : $E(\underline{\hat{b}}) = E((X'X)^{-1}X'\underline{y}) = (X'X)^{-1}X'E(\underline{y}) = (X'X)^{-1}X'X\underline{b} = \underline{b}$
- 추정분산 : $V(\underline{\hat{b}}) = V((X'X)^{-1}X'\underline{y}) = (X'X)^{-1}X')\sigma^2 I((X'X)^{-1}X')' = \sigma^2(X'X)^{-1}$

Gauss Marcov Theorem

OLS 추정치 $\underline{\hat{b}} = (X'X)^{-1}(X'\underline{y})$는 BLUE(Best Linear Unbiased Estimator)이다. [증명은 본 강의노트에 생략]

MLE 추정

$\underline{y} \sim N(X\underline{b}, \sigma^2 I_n)$이므로 우도함수는 $L(\underline{y}; X, \underline{b}, \sigma^2) \propto e^{-\frac{1}{2\sigma^2}(\underline{y}-X\underline{b})'(\underline{y}-X\underline{b})}$이므로

로그 우도함수를 최대화 하는 모수 $\underline{b}$ 추정치는 $\underset{\underline{\beta}}{min}(\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$을 최소화 하는 OLS 추정치와 동일하다. 그리고 MLE는 MVUE이므로 가장 좋은 추정치이다.

즉 OLS = MLE이다.

적합치와 잔차

적합치 fitted value $\hat{y}_i$

회귀모형에 OLS 추정치를 대체한 모형식 : $\underline{\hat{y}} = X\underline{\hat{b}} = X(X'X)^{-1}X'\underline{y} = H\underline{y}$

- 행렬 $H = X(X'X)^{-1}X'$ 를 hat 행렬이라 정의한다.
- H 행렬은 대칭행렬이며 멱등행렬이다. => 행렬 $(I - H)$도 대칭행렬이며 멱등행렬이다.

정의 : 만약 $AA = A$가 성립하는 행렬 A를 멱등행렬(idempodent)이라 한다. 만약 A가 멱등행렬이면, $A^n = A$(n은 2보다 큰 정수)가 성립한다.

잔차 residual 적합치와 관측치의 차이 $r_i = y_i - \hat{y}_i$

$\underline{r} = \underline{y} - \underline{\hat{y}} = \underline{y} - H\underline{y} = (I - H)\underline{y}$

- 잔차는 오차의 추정치이다. $\underline{\hat{e}} = \underline{r}$
- 잔차의 평균은 0이다. 분산은 $V(\underline{r}) = V((I - H)\underline{y}) = \sigma^2(I - H)$

OLS 추정치 성질

GAUSS-MARKOV Theorem : 회귀계수에 대한 OLS 추정치는 BLUE(Best Linear Unbiased Estimator)이다. 즉 모든 선형, 불편 추정량 중 최소 분산(minimum variance)를 갖는다. 분포함수가 지수족이므로 MLE는 CSS이고 불편추정량이므로 Rao-Blackwell 정리에 의해 MVUE이다.

[증명] 행렬로 증명하는 것이 간편하고 다중회귀모형도 동일하게 적용할 수 있어 행렬로 증명함.

행렬 모형 : $\underset{\sim}{y}_{(n\times 1)} = X_{(n\times(p+1))}\underset{\sim}{b}_{((p+1)\times 1)} + \underset{\sim}{e}_{(n\times 1)},\ \underset{\sim}{e} \sim N(\underset{\sim}{0}, \sigma^2 I)$

설명변수 개수 p, 데이터 크기 n 인 모형

OLS : $min_{\underset{\sim}{b}}\ (\underset{\sim}{e}'\underset{\sim}{e}) => \underset{\sim}{\hat{b}} = (X'X)^{-1}(X'\underset{\sim}{y})$: $\underset{\sim}{\hat{b}} = K_{(p+1)\times n}\underset{\sim}{y}$: OLS 추정치는 $\underset{\sim}{y}$ 선형함수

또다른 선형 추정치를 $\underset{\sim}{\hat{b}}^{*} = (K+C)\underset{\sim}{y}$ 라 하자. [궁극적으로 C은 0 행렬임을 보이면 된다]

불편성 : $E(\underset{\sim}{\hat{b}}) = E(K+C)\underset{\sim}{y}) = E(K\underset{\sim}{y}) + E(C\underset{\sim}{y}) = \underset{\sim}{b} + CX\underset{\sim}{b}$ 그러므로 $CX = \underset{\sim}{0}$이어야 $\underset{\sim}{b}^{*}$가 불편성을 만족한다. 아직 C=0행렬을 보인 것은 아니다. 단지 $CX = \underset{\sim}{0}$임을 보였을 뿐이다.

최소분산성 :

$V(\underset{\sim}{b}^{*}) = V[(K+C)\underset{\sim}{y}] = (K+C)V(\underset{\sim}{y})(K+C)' = \sigma^2(K+C)(K+C)'$

정리하면 $V(\underset{\sim}{b}^{*}) = \sigma^2(KK'+CC') = V(\underset{\sim}{\hat{b}}) + \sigma^2 CC'$

CC'는 양반정치행렬(positive semi-definite matrix)이므로 음이 될 수 없으므로 OLS 추정치 $\underset{\sim}{\hat{b}}$의 추정분산이 불편, 선형 추정치 중 최소 분산이다. **(QED)**

Inference

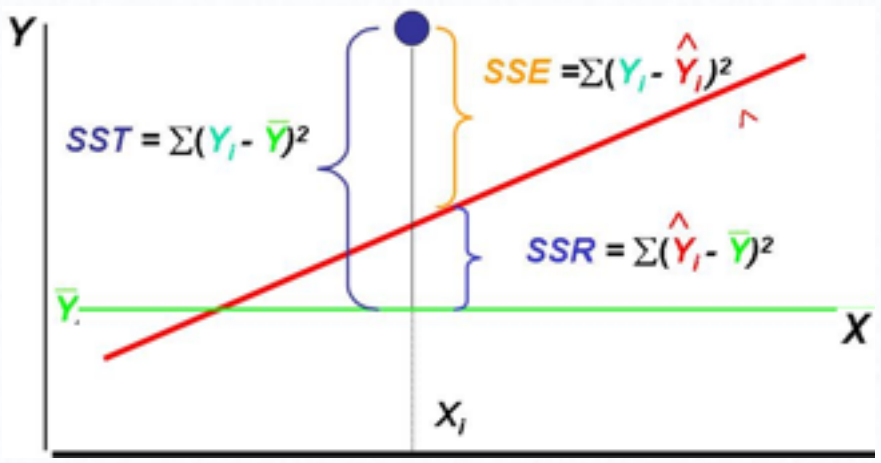
분산분석적 접근

개념

목표변수의 분산(변동)을 모형설명변동과 설명하지 못하는 변동으로 나누어 non-설명변동 대비 설명변동이 충분히 크면 모형이 유의하다고 판단한다.

변동은 데이터 값의 변화에 대한 측정이므로 데이터의 정보와 동일함

정규분포를 따르는 확률변수(목표변수 y_i가 이에 해당)의 변동(분산 계산과 동일)은 카이제곱 분포를 따르고 변동의 비 (카이제곱분포의 비)는 F-분포를 따르므로 이를 이용하여 모형의 유의성을 검증



변동분할 variation decomposition

총변동 Total Sum of Square $SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n = \underline{y}'\underline{y} - (1/n)\underline{y}'J_n\underline{y}$

- 목표변수의 값들의 변동에 대한 척도로 예측변수들이 목표변수를 잘 설명한다는 것은 변화를 예측하는 것임 = 목표변수의 분산과 동일 개념
- Quardratic form(이차형식) : $SST = \underline{y}'(I - (1/n)J)\underline{y}$

오차변동 Error SS $SSE = \sum (y_i - \hat{y})^2 = (\underline{y} - H\underline{y})'(\underline{y} - H\underline{y}) = \underline{y}'(I - H)\underline{y}$

- 총변동 중 회귀모형에 의해 설명되지 못하는 부분

모형변동 Regression (Model) SS $SSE = \sum (\hat{y}_i - \bar{y})^2 = \underline{y}'(H - (1/n)J)\underline{y}$

- 목표변수 변동 중 설정된 예측변수들에 의해 설명되어지는 변동
- 총변동에 가까울수록 회귀모형 설정이 적절함을 의미함

Quardratic Form : 벡터 $\underline{y}$에 대하여 $Q = \underline{y}'A\underline{y}$를 이차형식이라 한다.

[이차형식 분포 정리] 만약 $\underline{y} \sim MN(,)$(정규분포를 따른다면), 이차형식은 $Q = \underline{y}'A\underline{y}$는 자유도가 rank(A)인 카이제곱분포를 따른다.

만약 $\underline{y} \sim MN(\underline{\mu}, \sigma^2 I)$이면, 평균 $E(Q) = \underline{y}'\underline{\mu}\underline{y}$, 공분산 $COV(Q) = A\sigma^2 A'$

변동의 분포

오차변동 분포 : $SSE \sim \chi^2(n-p-1)$

- $SSE = \underline{y}'(I-H)\underline{y}$ 는 목표변수 $\underline{y}$의 선형함수이다.
- 오차항의 정규성 가정에 의하여 $\underline{y} \sim MN(X\underline{b}, \sigma^2 I)$ 정규분포를 따르므로 SSE도 정규분포를 따른다.
- 행렬 (I-H)의 계수(rank)가 (n-p-1)이다.

총변동 분포 : $SST \sim \chi^2(n-1)$

- $SST = \underline{y}'(I-(1/n)J)\underline{y}$ 이차형식이고 $(I-(1/n)J)$ 자유도가 (n-1)이다.

자유도 분할: Cochran 정리

- 총변동의 자유도(관측치 중 자유로운 개수, 관측치 하나 하나는 독립적이고 정보를 갖고 있다)는 평균이 추정되었으므로 (n-1)이다.
- SSE의 자유도는 (n-1-p)이다. 왜냐하면 p개의 모수가 추정되었기 때문이다.
- 그러므로 SSR의 자유도는 SST 자유도로부터 SSE 자유도를 뺀 값으로 p이다.

총변동 분포 : $SSR \sim \chi^2(p)$

오차항 분산(σ^2) 추정

$$\hat{\sigma^2} = SSE/(n-p-1) = MSE$$

분산분석표

변동(source)	SS(자승합)	자유도	MS(평균자승합)	F-검정
Regression (모형)	$SSR = \underline{y}'[H - (\frac{1}{n})J]\underline{y}$	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error (오차)	$SSE = \underline{y}'[I - H]\underline{y}$	n-p-1	$MSE = \frac{SSE}{n-2}$	$\sim F(p, n-p-1)$
Total (총 변동)	$SST = \underline{y}'[I - (\frac{1}{n})J]\underline{y}$	n-1	결정계수: $R^2 = \frac{SSR}{SST}$	

F - 검정

- 귀무가설 : 설정한 회귀모형은 유의하지 않다. <=> 회귀계수 벡터 $\underline{b}$ 모든 계수는 0이다.
- 대립가설 : 설정한 회귀모형은 유의하다. <=> 회귀계수 벡터 $\underline{b}$ 중 유의한 회귀계수가 적어도 하나가 있다.
- 그러므로 F-검정 결과 귀무가설이 기각되면 유의한 설명 변수가 하나 이상 있다는 것이므로 각 설명 변수에 대한 유의성을 t-검정을 이용하여 알아 보면 된다.

결정계수 : $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

수정된 결정계수 : $R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$

파이썬 코드 (statsmodel 모듈 이용)

기초통계량

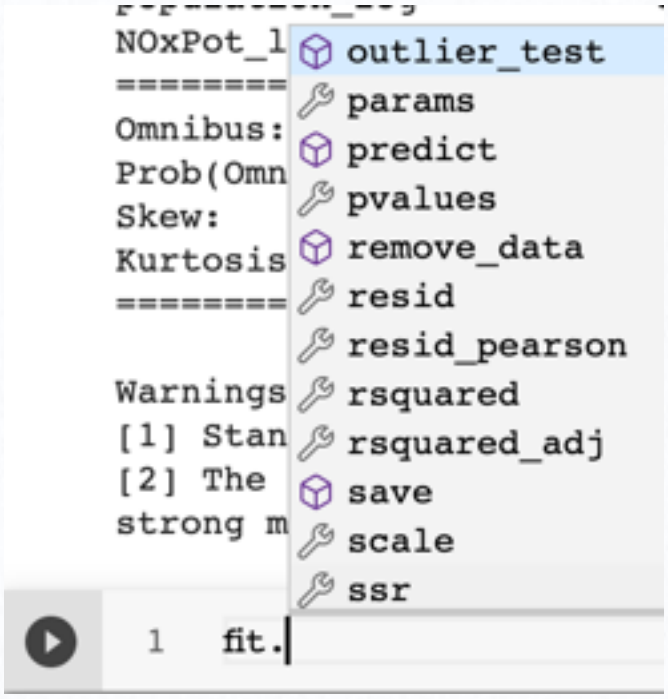
```
smsa_clean.describe()
```

```
smsa_clean.groupby('southern').describe()
```

	jan_temp	july_temp
count	55.000000	55.000000
mean	34.163636	74.545455
std	10.418580	4.729021
min	12.000000	63.000000
25%	27.000000	72.000000
50%	32.000000	74.000000
75%	40.000000	77.500000
max	67.000000	85.000000

	jan_temp		
	count	mean	std
southern			
NO	42.0	30.809524	8.399298
YES	13.0	45.000000	8.990736

```
import statsmodels.api as sm
y=smsa_clean['mortality_index']
X=sm.add_constant(smsa_clean.drop(columns=['mortality_index','southern'])) #add intercept
fit=sm.OLS(y, X).fit()
print(fit.summary())
```



sm.OLS() 실행하여 얻은 결과는 모두 fit에 저장되어 있으며 어떤 통계량(결과)들이 저장되어 있는지는 fit.을 입력하면 자동 팝업된다.

결정계수 78.1%로 설정된 모형이 사망률 지수(목표변수)를 충분히 설명하고 있음

```
[144]  1  print('결정계수=%.2f(%%)' % (fit.rsquared*100))
```

```
결정계수=77.34(%)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        mortality_index   R-squared:                       0.773
Model:                            OLS   Adj. R-squared:                  0.701
Method:                 Least Squares   F-statistic:                     10.73
Date:                Sat, 18 Apr 2020   Prob (F-statistic):           5.70e-10
Time:                        01:28:24   Log-Likelihood:                -283.32
No. Observations:                  59   AIC:                             596.6
Df Residuals:                      44   BIC:                             627.8
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1247.4636    334.995      3.724      0.001     572.326    1922.601
jan_temp         -1.7404      0.750     -2.322      0.025      -3.251      -0.230
july_temp        -1.6887      1.952     -0.865      0.392      -5.622       2.244
humidity         -0.5929      1.153     -0.514      0.610      -2.917       1.731
rainfall          1.6976      0.558      3.045      0.004       0.574       2.821
education       -12.1181      8.494     -1.427      0.161     -29.237       5.001
white_color_ratio  -1.7787      1.139     -1.562      0.125      -4.074       0.516
person_houshold  -51.4818     40.907     -1.259      0.215    -133.924      30.960
household_income   -0.0006      0.001     -0.434      0.667      -0.003       0.002
HCPot              -0.1641      0.096     -1.716      0.093      -0.357       0.029
SO2Pot              0.0006      0.120      0.005      0.996      -0.241       0.242
pop_density_log     5.8923     16.400      0.359      0.721     -27.160      38.944
non_white_ratio_sqrt  32.1485      6.255      5.140      0.000      19.543      44.754
population_log      3.8755      7.546      0.514      0.610     -11.332      19.083
NOxPot_log         21.0724      8.567      2.460      0.018       3.808      38.337
==============================================================================
```

설정된 모형이 사망률 지수를 충분히 설명한다는 것이 18개 예측변수 모두 유의한 설명을 하고 있다는 것과 동일한 것은 아니다. 각 예측변수의 유의성은 t-분포 검정통계량에 의해 검정되며 유의확률에 의해 판단된다. 1월 기온(-), 강우량(+),… 유의하다. 비백인비율-제곱근 변환, NOxPot_log 매우 유의

```
1  print('회귀계수 : 유의확률\n',fit.pvalues)
```

```
회귀계수 : 유의확률
 const                0.000555
jan_temp              0.024953
july_temp             0.391556
humidity              0.609747
rainfall              0.003924
education             0.160742
white_color_ratio     0.125447
```

분산분석표 및 결정계수

```
print('모형변동=%.2f | 오차변동=%.2f | 총변동=%.2f'%
(fit.mse_model*(X.shape[1]-1),fit.mse_resid*(fit.nobs-X.shape[1]-1),fit.mse_total*(fit.nobs-1)))
```

```
모형변동=174785.74 | 오차변동=50043.00 | 총변동=225992.53
```

예측치 및 예측구간

```
fit.fittedvalues #적합치
```

```
state_nm   city_nm
 OH        Akron                    948.902228
 NY        Albany-Schenectady-Troy  921.123077
 PA-NJ     Allentown                919.265882
 GA        Atlanta                  973.169818
```

```
from statsmodels.sandbox.regression.predstd import wls_prediction_std
prstd, iv_l, iv_u=wls_prediction_std(fit) #추정오차, 예측구간 하한, 상한
```

lm=LinearRegression() 이용

- intercept 항은 자동 삽입되므로 sm 모듈에서 add 절편 과정은 필요 없다.
- 그러나 저장되는 통계량이 적어 자주 사용하지 않는다.

```
from sklearn.linear_model import LinearRegression
lm=LinearRegression()
y=smsa_clean['mortality_index']
X=smsa_clean.drop(columns=['mortality_index','southern'])
lm.fit(X,y)
params = np.append(lm.intercept_,lm.coef_) #OLS estimates
fitted=lm.predict(X)   #fitted value
r2=lm.score(X,y) #R-square; print(params, r2)
```

```
[1435.9895   -1.6744   -2.0509   -0.4938    1.4035  -12.1316    0.0022
    2.2291   -1.9583    0.      -55.4049    0.0002   -0.6007    0.8728
   -0.0583   -0.2761   18.1517   -3.1259   18.0095] 0.7806036890032
```